Vergleich von Strategien zum Clustern von Daten mit fehlenden Werten

Ludmila Himmelspach
Institut für Informatik
Heinrich-Heine-Universität Düsseldorf
D-40225 Düsseldorf, Deutschland
himmelspach@cs.uni-duesseldorf.de

Zusammenfassung

Die klassischen Methoden zur Clusteranalyse wurden entwickelt um auf vollständigen Daten Analysen durchzuführen. Oft fehlen aber in Daten einzelne Werte — systematisch oder unsystematisch —, z.B. infolge der Probleme bei der Datenerfassung, Datenübertragung, Datenbereinigung oder weil Daten aus unterschiedlichen Quellen stammen. Demzufolge können die traditionellen Clusteringmethoden zur Analyse solcher Daten nicht ohne weiteres angewendet werden. Im Rahmen dieses Beitrags werden unterschiedliche Strategien zum Umgang mit fehlenden Werten in Daten für das Clusteringproblem vorgestellt, analysiert und miteinander verglichen. Dabei wird das besondere Augenmerk auf die Untersuchung der Leistungsfähigkeit dieser Verfahren in Abhängigkeit von den Ausfallmechanismen, die den fehlenden Werten zugrundeliegen, und von dem Anteil fehlender Werte in Daten gelegt.

1 Einleitung

Mit dem rasanten Anstieg an Möglichkeiten große Datenmengen elektronisch zu erfassen und zu speichern, haben auch Werkzeuge zur Datenanalyse stark an Bedeutung gewonnen. Diese großen Mengen von Daten können viel potentiell wichtiges Wissen enthalten, das aber zuerst im Rahmen eines *Knowledge Discovery in Databases*-Prozess aus den Daten extrahiert werden muss. Bei der Analyse der Daten geht es oft im ersten Schritt darum auf der riesigen Datenmenge Gruppen von ähnlichen Objekten zu identifizieren, was die Aufgabe der Clusteranalyse ist. Die Methoden der Clusteranalyse finden in vielen Bereichen ihre Anwendung, einschließlich Database Marketing, Web-Analyse, Information Retrieval, Bioinformatik und weiteren.

Den Ausgangspunkt für die Clusteranalyse bilden Daten. Dabei handelt es sich um eine Menge von Merkmalswerten, die in Form einer Datenmatrix für die Auswertung vorliegen. Oft kommen aber in Daten fehlende Werte vor, die z.B. durch Fehler bei der Datenerfassung, Datenübertragung oder Datenbereinigung verursacht werden konnten. Die fehlenden Werte können einer zufälligen Anordnung oder nach bestimmten Mustern in Datenmatrizen vorkommen. Sie können zufälligen oder systematischen Ausfallmechanismen unterliegen.

Die klassischen Methoden zur Clusteranalyse wurden entwickelt, um auf vollständigen Datenmatrizen Analysen durchzuführen. In den Fällen, wo die Vervollständigung der Datensätze durch Datennacherhebung z.B. aus Kosten- oder Zeitgründen unerwünscht oder sogar unmöglich ist, braucht man Analysemethoden, die mit dem Problem fehlender Werte in Daten umgehen können. Zur Behandlung von Daten, die fehlende Werte beinhalten, gibt es im Allgemeinen drei verschiedene Ansätze [LR02, Wag04]. Der erste Ansatz basiert auf der Eliminierung der Datensätze bzw. Merkmale, die fehlende Werte aufweisen. Beim zweiten Ansatz werden fehlende Werte im Rahmen einer Datenvorverarbeitung geschätzt. Der dritte Ansatz besteht darin, die datenanalytischen Verfahren für den Umgang mit fehlenden Werten zu adaptieren. Im Rahmen dieses Beitrags werden diese drei Strategien zum Umgang mit fehlenden Werten in Daten

für das Clusteringproblem am Beispiel des Fuzzy C-Means-Algorithmus anhand eines geeigneten Datensatzes analysiert und miteinander verglichen. Dabei untersuchen wir insbesondere die Leistungsfähigkeit dieser Verfahren in Abhängigkeit von den Ausfallmechanismen, die den fehlenden Werten zugrundeliegen, und von dem Anteil fehlender Werte im Datensatz.

2 Grundlagen

2.1 Fuzzy C-Means-Algorithmus

Der *Fuzzy C-Means-Algorithmus (FCM)* gehört zu den partitionierenden Clusteringalgorithmen, d.h. die zu klassifizierende Datenmenge wird vollständig in eine vorgegebene Anzahl von Cluster zerlegt. Im Unterschied zu den klassischen partitionierenden Clusteringmethoden wird die Zuordnung der Datenpunkte zu den Clustern bei FCM durch die Zugehörigkeitsgrade ausgedrückt [Bez81]. Der Zugehörigkeitsgrad eines Objekts bezüglich eines Clusters drückt aus, wie sicher dieses Objekt dem Cluster zuzuordnen ist. Die Zugehörigkeitsgrade werden basierend auf dem Abstand der Datenpunkte zu den Clustern berechnet und liegen im Intervall zwischen 0 und 1. Dabei zeigt 0 keine Zugehörigkeit des Objektes zu dem betreffenden Cluster. Ein Zugehörigkeitsgrad von 1 zeigt an, dass das Objekt dem Cluster mit Sicherheit zuzuordnen ist.

Wie die meisten partitionierenden Clusteringverfahren findet auch FCM eine optimale Zerlegung der Datenmenge durch die Minimierung der Zielfunktion. Die Zielfunktion für FCM berechnet für alle Cluster die Summen der quadrierten und durch die Zugehörigkeitsgrade gewichteten Abstände der Datenpunkte zu den jeweiligen Clusterzentren und addiert diese Teilsummen. Das heißt, die Cluster sollen so gebildet werden, dass die Abstände der Datenpunkte zu den Clusterzentren minimal sind. Da die Zielfunktion nicht direkt optimiert werden kann, wird sie in jedem Iterationsschritt von FCM bezüglich der Zugehörigkeitsgrade und Clusterzentren minimiert.

2.2 Arten von Ausfallmechanismen

Ein wichtiger Faktor bei der Auswahl eines datenanalytischen Verfahrens, das mit fehlenden Werten in Daten umgehen kann, ist der zugrundeliegende Mechanismus, der zum Ausfall von Daten geführt hat. Neben dem zufälligen Fehlen von Werten in Datenmatrizen kann es sein, dass das Fehlen eines Wertes von der Ausprägung seines Attributs oder von den Ausprägungen anderer Attribute abhängt. Grundsätzlich werden in der Literatur zwei Arten von Ausfallmechanismen unterschieden: unsystematischer (d.h. zufällig fehlend) und systematischer (d.h. nicht zufällig fehlend) Ausfallmechanismus [Ban95]. Der systematische Ausfallmechanismus liegt vor, wenn das Fehlen der Werte von der Ausprägung des Merkmals selbst abhängt, in dem sie fehlen. Die fehlenden Daten werden dann als *„not missing at random“ (NMAR)* bezeichnet [LR02]. Der unsystematische Ausfallmechanismus kann zusätzlich in zwei Klassen eingeteilt werden: *„missing at random“* und *„missing completely at random“*. Die fehlenden Werte in der Datenmatrix werden als *„missing at random“ (MAR)* bezeichnet, wenn das Fehlen der Werte allein von den Ausprägungen der beobachteten Merkmale abhängt. Wenn das Fehlen der Werte in der Datenmatrix unabhängig von Ausprägungen der Attribute ist, unabhängig davon, ob sie beobachtet wurden oder fehlen, dann spricht man von fehlenden Werten *„missing completely at random“ (MCAR).*

3 Strategien zum Umgang mit fehlenden Werten

3.1 Eliminierungsverfahren

Die einfachste Methode mit unvollständigen Daten umzugehen ist die Datensätze oder Merkmale mit fehlenden Werten aus der Datenmatrix zu eliminieren und die Datenanalyse nur auf Grund der vollständig erhobenen Datenobjekte bzw. Merkmale durchzuführen [LR02, Ban95].

Werden Datensätze mit fehlenden Werten von der Analyse ausgeschlossen, wird dieses Verfahren in der Literatur als *„complete-case analysis“* bezeichnet. Werden Merkmale, in denen Datensätze fehlende Werte aufweisen, bei der Datenanalyse nicht betrachtet, so wird dieses Verfahren als *„compelete-variable analysis“* bezeichnet. Die erste Methode wird im Allgemeinen dann angewendet, wenn der Anteil der Datenobjekte mit fehlenden Werten relativ gering ist und bei der Datenanalyse nicht alle Datenobjekte berücksichtigt werden müssen. Ist der Anteil der Datensätze mit fehlenden Werten hoch oder müssen alle Datensätze klassifiziert werden, so eignet sich die zweite Methode besser, wobei auch hier der Anteil der Merkmale mit fehlenden Werten nicht zu hoch sein darf, da sonst diese Vorgehensweise zum Verlust einer für das Clustering aussagekräftigen Dimension führen kann. Trotz der Nachteile wird dieses Verfahren häufig als Default-Ansatz implementiert und als Maßstab für andere Verfahren verwendet.

3.2 Imputationverfahren

Eine andere Möglichkeit mit fehlenden Werten in Datenmatrizen bei der Datenanalyse umzugehen ist die fehlenden Werte im Rahmen einer Datenvorverarbeitung zu schätzen. In der Literatur wird für diese Methode der Begriff *„missing value imputation“* verwendet. Neben der Imputation durch zufällige Auswahl der vorhandenen Werte aus der Datenmatrix, gibt es zahlreiche statistische Verfahren, um fehlende Werte zu schätzen. Die Imputationstechniken reichen von den einfachsten wie z.B. Ergänzung der unvollständigen Datenmatrix durch Minimum, Maximum oder Mittelwert vorhandener Werte, bis zu statistischen Verfahren wie z.B. mittels Regressions-, Varianz- oder Hauptkomponentenanalyse, die versuchen die Zusammenhänge zwischen den Attributen aufzudecken und diese dann zur Bestimmung der Imputationswerte zu nutzen [LR02]. Auch der Expectation-Maximization-Algorithmus (EM) wird oft zur Schätzung fehlender Werte verwendet [DLR77]. Der Hauptvorteil der Imputationsverfahren liegt darin, dass die anschließende Datenanalyse auf der vollständigen Datenmatrix wie im Fall ohne fehlende Werte erfolgen kann. Der Nachteil dieser Vorgehensweise ist neben dem hohen Rechenaufwand jedoch, dass die Ergebnisse der Datenanalyse durch die verwendeten Imputationstechniken stark beeinflußt werden, da während der Datenanalyse zwischen den beobachteten und geschätzten Werten nicht mehr unterschieden wird.

3.3 Adaptierte Clusteringverfahren für Daten mit fehlenden Werten

Der letzte Ansatz für den Umgang mit fehlenden Werten in Daten ist die datenanalytischen Verfahren so zu ändern, dass diese bei der Analyse die Datenobjekte mit fehlenden Werten im vollen Umfang berücksichtigen. Die Strategien den Fuzzy C-Means-Algorithmus an Daten mit fehlenden Werten zu adaptieren, kann man im Allgemeinen in zwei Kategorien unterteilen. Zur ersten Kategorie gehören Verfahren, die beim Clustern unvollständiger Datensätze nur vorhandene Werte einbeziehen. Zur zweiten Kategorie gehören Verfahren, die beim Clustern in jedem Iterationsschritt fehlende Werte in Abhängigkeit von den Clusterzentren oder vorhandenen Attributwerten schätzen und ersetzen. Wir beschränken unsere Betrachtung hier auf drei Verfahren, wobei die ersten beiden zu der ersten und das dritte zu der zweiten Kategorie gehören.

Whole-data strategy (WDS): Bei dieser Methode werden zuerst alle vollständig vorhandenen Datensätze mit FCM klassifiziert. Danach werden die Daten mit fehlenden Werten unter Berechnung der partiellen Distanzen jeweils dem nächstliegenden Clusterzentrum zugeordnet [HB01].

Partial distance strategy (PDS): Dieses Verfahren verwendet partielle Distanzen zwischen den Datenpunkten mit fehlenden Werten [HB01].

Nearest prototype strategy (NPS): Die fehlenden Attributwerte eines Datenpunktes werden durch die entsprechenden Werte des nächstliegenden Clusterzentrums in jedem Itera-

tionsschritt ersetzt [HB01]. Bei der Distanzberechnung wird dabei die partielle Distanzfunktion verwendet [Dix79].

Eine Übersicht über weitere Strategien und sowie deren Vergleich kann in [Him08] gefunden werden.

4 Datenexperimente und Ergebnisse

Die oben beschriebenen Verfahren zum Umgang mit fehlenden Werten wurden am Beispiel von FCM anhand eines künstlichen Datensatzes untersucht, der durch eine Mischung von drei 3-dimensionalen Gaußverteilungen generiert wurde. Der Datensatz besteht aus 100 Datenpunkten, wobei sich diese gleichmäßig auf drei Cluster verteilen (siehe Abbildung 1). Da die abhängigen Dimensionen für das Clustering keine zusätzlichen Informationen liefern, wurden die Daten so generiert, dass es keine Abhängigkeiten zwischen den Werten verschiedener Dimensionen gibt.

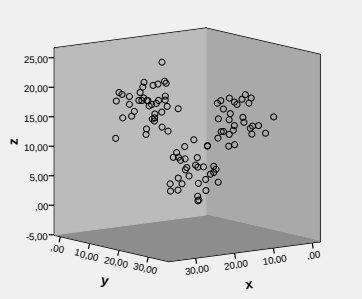


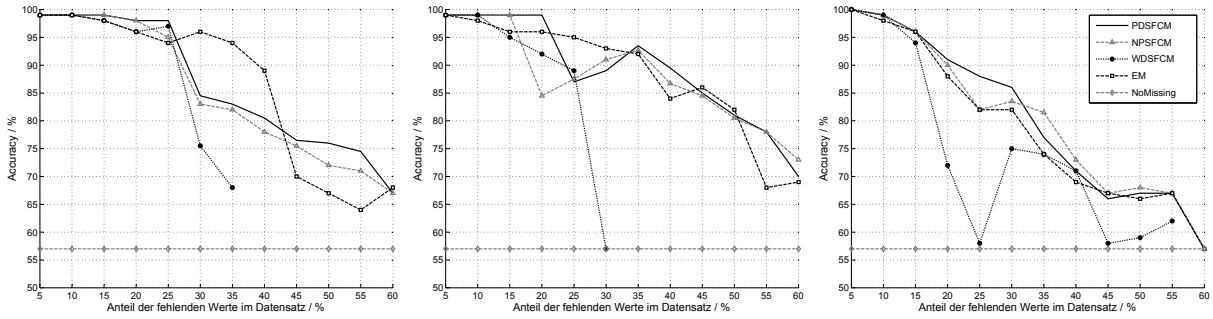
Abbildung 1: Testdatensatz

Um die im Abschnitt 3 vorgestellten Verfahren hinsichtlich ihrer Leistungsfähigkeit anhand der Daten testen zu können, wurden aus dem Datensatz Werte in zwei der drei Dimensionen schrittweise entfernt, wobei sich der Anteil fehlender Werte auf die Gesamtanzahl der Werte im Datensatz bezieht. Um auch prüfen zu können, ob die Leistungsfähigkeit der einzelnen Algorithmen von den unterschiedlichen Ausfallmechanismen abhängt, wurden die Werte gemäß der im Abschnitt 2.2 beschriebenen Ausfallmechanismen MCAR, MAR und NMAR aus dem Datensatz entfernt.

Abbildung 2 zeigt die Ergebnisse der Performance-Analyse der Verfahren PDSFCM, NPSFCM, WDSFCM, EM und NoMissing in Abhängigkeit von dem Anteil fehlender Werte MCAR, MAR und NMAR für die Attribute y und z. Bei den ersten drei Verfahren handelt es sich um die Algorithmen aus dem Abschnitt 3.3. Im Rahmen des EM-Verfahrens wurden fehlende Werte mittels EM-Algorithmus (vgl. Abschnitt 3.2) geschätzt und anschließend mit FCM klassifiziert. Das Verfahren NoMissing entspricht dem „complete-variable“-Ansatz aus Abschnitt 3.1. Als Bewertungsmaß für die Algorithmen wurde der über 10 Durchläufe gemittelte Wert für die Accuracy verwendet, der den Anteil richtig klassifizierter Datenpunkte zur Gesamtanzahl der Objekte im Datensatz prozentual angibt.

Wie man in den Diagrammen erkennen kann, unterscheiden sich die Algorithmen PDSFCM, NPSFCM, WDSFCM und EM hinsichtlich ihrer Leistungsfähigkeit nur unwesentlich voneinander und liegen weit über den Ergebnissen der anderen zwei Verfahren. Die Accuracy-Werte für diese Algorithmen liegen bei einem kleinen Anteil (bis 20%) fehlender Werte im Datensatz für alle Ausfallmechanismen über 90%. Mit steigender Anzahl fehlender Werte im Datensatz unterscheidet sich die Leistungsfähigkeit der Algorithmen in Abhängigkeit von den zugrundeliegenden Ausfallmechanismen. Die besten Ergebnisse sind beim Ausfallmechanismus MAR zu beobachten. Wenn hingegen der Ausfallmechanismus NMAR vorliegt, fallen die Ergebnisse der Algorithmen am schlechtesten aus. Die guten Ergebnisse beim Vorliegen des Ausfallmechanismus MAR sind unter anderem dadurch zu erklären, dass der Anteil der Datensätze mit zwei fehlenden Werten in Vergleich zu anderen Ausfallmechanismen sehr klein ist. Andere Experimente haben jedoch gezeigt, dass die Leistungsfähigkeit der Algorithmen auf Daten mit fehlenden Werten MAR mit hohem Anteil der Datensätze mit zwei fehlenden Werten viel schlechter ist als z.B. beim Vorliegen des Ausfallmechanismus MCAR. Bei einem hohen Anteil fehlender Werte lag die Accuracy für diese Algorithmen sogar unter der für das NoMissing-Verfahren (vgl. [Him08]).

Wie Diagramme in Abbildung 2 zeigen, konnte die Accuracy für den Algorithmus WDSFCM nicht immer berechnet werden. Das liegt daran, dass es ab einem bestimmten Anteil fehlender Werte im Datensatz keine vollständigen Datenobjekte mehr gab, was den Einsatz von WDSFCM



(a) (b) (c)

Abbildung 2: Accuracy für verschiedene Algorithmen in Abhängigkeit von dem Anteil fehlender Werte (a) MCAR, (b) MAR und (c) NMAR für die Attribute y und z.

unmöglich machte. Da dieser Algorithmus Clusterzentren ausschließlich anhand der vollständigen Datensätze berechnet, hängt die Verteilung der Datenpunkte auf Cluster sehr stark von der Verteilung der vollständigen Datensätze ab, was die schlechten Ergebnisse des Algorithmus bei einem hohen Anteil fehlender Werte im Datensatz erklärt.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurden unterschiedliche Strategien zum Umgang mit fehlenden Werten in Daten für das Clusteringproblem anhand eines künstlichen Datensatzes analysiert und miteinander verglichen. Die Testergebnisse haben gezeigt, dass es sinnvoll ist bei der Clusteranalyse alle vorhandenen Werte zu berücksichtigen. So haben die Imputations- und adaptierte Clusteringverfahren bei Experimenten viel bessere Ergebnisse erzielt als das Eliminierungsverfahren. Außerdem haben wir gezeigt, dass die Qualität der Ergebnisse stark von dem Ausfallmechanismus abhängt, der den fehlenden Werten zu Grunde liegt. Deswegen wird das nächste Forschungsziel sein die Verfahren für unterschiedliche Ausfallmechanismen anzupassen, um dadurch bessere Ergebnisse zu erzielen.

Literatur

[Ban95] U. Bankhofer. *Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse.* Eul, Bergisch-Gladbach, 1995.

[Bez81] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers, 1981.

[Dix79] J.K. Dixon. Pattern Recognition with Partly Missing Data. *IEEE Transactions on System, Man and Cybernetics*, 9:617–621, 1979.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelyhood from Incomplete Data via EM Algorithm. *Journal of the Royal Statistical Society, Series B*, pages 1–31, 1977.

[HB01] R.J. Hathaway and J.C. Bezdek. Fuzzy c-means Clustering of Incomplete Data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 31(5):735–744, 2001.

[Him08] L. Himmelspach. Clustering mit fehlenden Werten: Analyse und Vergleich. Masterarbeit, Institut für Informatik, Heinrich-Heine-Universität Düsseldorf, 2008.

[LR02] R.J. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, 2002.

[Wag04] K. Wagstaff. Clustering with Missing Values: No Imputation Required. In *Classification, Clustering, and Data Mining Applications (Proceedings Meeting of the International Federation of Classification Societies)*, pages 649–658, 2004.